

Stat 321 - Day 29
Descriptive Statistics

Today we take a break from studying probability and instead shift gears to examine procedures for analyzing data descriptively. A common theme will be to start with *graphical displays* of data and then proceed to calculate *numerical summaries* of the *center* and *spread* of the data.

Example: Rowers Weights

The following data are the weights (in pounds) of the 27 members of the U.S. men's rowing team at the 2004 Olympics (also found in the Minitab worksheet `rowers04.mtw`):

Name	Wgt	event	name	wgt	event	name	wgt	event
Abdullah	185	double sculls	Ruckman	155	LW double sculls	Smith	160	LW four
Samsonov	185	pair	Nuzum	210	double sculls	Wherley	215	four
Hoopman	185	eight	DuRoss	210	quad sculls	Todd	154	LW four
Holbrook	195	quad sculls	Schroeder	229	four	Teti	160	LW four
Wilkinson	190	quad sculls	Read	180	eight	Cipollone	120	eight
Volpenhein	215	eight	Hansen	210	eight	Tucker	153	LW double sculls
Ahrens	215	eight	Smack	195	quad sculls	Warner	160	LW four
Beery	215	eight	Walton	180	pair	Moser	210	four
Klugh	205	four	Deakin	200	eight	Allen	210	eight

- (a) Use Minitab to produce a *dotplot* of the rowers' weights (choose `Graph > Dotplot` from the menu, or type `MTB > dotplot c2`). Comment on what the graph reveals about the distribution of weights among the rowing team members. Specifically, comment on the *center* of the distribution, on how *spread* out the data are, on the *shape* of the distribution, and on *unusual features* or observations. [*Hints*: Imagine that you are describing these weights to someone who knows absolutely nothing about rowing or even about what people weigh. Be sure to relate your comments to the context.]
- (b) How many groups/clusters do you see in the dotplot? Suggest an explanation, based on the context, for these groups. [*Hint*: Look at the event names.]
- (c) Use Minitab to produce a histogram of the rowers' weights (choose `Graph > Histogram` from the menu, or type `MTB > histogram c2`). Does the histogram reveal the three groups as clearly as the dotplot did?

- (d) Use Minitab to change the number of intervals in the histogram (MTB> hist c2; SUBC> nint 10). Then try several choices for the number of intervals. Report the number that you think provides the most informative display of the distribution.

A third visual display is called a *stemplot* (or stem-and-leaf plot). Such a plot breaks up each value into a stem piece and a leaf piece. It then displays the stems in a vertical list, with leaves ordered left-to-right on the row with their stems.

- (e) The following stemplot includes only the rowers in the first column of the table above. Fill in the remainder of the stemplot.

12	
13	
14	
15	
16	
17	
18	555
19	05
20	5
21	555
22	

- (f) Is the distribution of rower weights symmetric, skewed to the right, or skewed to the left?

Two common measures of the center of a distribution are:

- *Mean*: arithmetic average
- *Median*: middle value, found in position $(n+1)/2$ if there are an odd number of observations, and defined to be the average of the two middle values if there are an even number of observations

- (g) Use the stemplot to determine the median of the rower weights (by hand).
- (h) Use Minitab to calculate the mean of the rower weights and to confirm your median calculation (MTB> desc c2).

Mean:

Median:

- (i) How do the mean and median compare to each other? Explain why this makes sense considering the shape of the distribution.

(j) If Cipollone is removed from the list, how would you expect the mean or median to change? Explain.

(k) Delete Cipollone's weight in the Minitab worksheet (just highlight his weight and hit the backspace key). Re-calculate the mean and median. How did they change? Which changed more substantially?

Mean:

Median:

(l) Now delete all of the lightweight rowers' weights in addition to Cipollone's. Re-calculate the mean and median. How did they change? Which changed more substantially?

Mean:

Median:

(m) Now take the heaviest rower and add 100 pounds to his weight. Re-calculate the mean and median. How did they change? Which changed more substantially?

Mean:

Median:

A statistic that is not substantially affected by extreme observations is said to be *resistant to outliers*.

(n) Which measure of center (mean or median) is resistant to outliers?

Two common measures of the spread, or variability, of a distribution are:

- *Standard deviation*: $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$, where \bar{x} denotes the mean
- *Inter-quartile range* (IQR, also known as fourth spread): the difference between the upper and lower quartiles, also known as upper and lower fourths, or as 75th and 25th percentiles. The upper quartile is calculated as the median of the values above the location of the overall median, and the lower quartile is calculated as the median of the values below the location of the overall median. Be aware that Minitab uses a slightly more complicated algorithm for calculating quartiles.

(o) Use the stemplot above to determine the quartiles and then the IQR.

Upper quartile:

Lower quartile:

IQR:

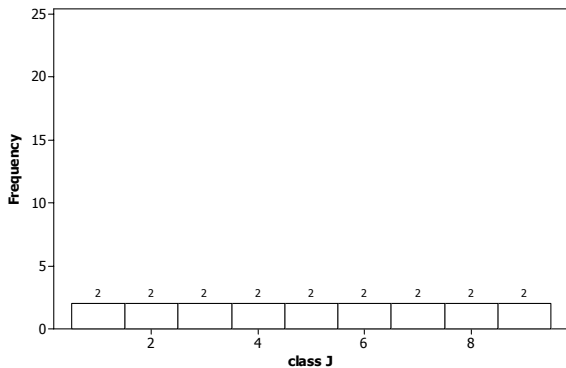
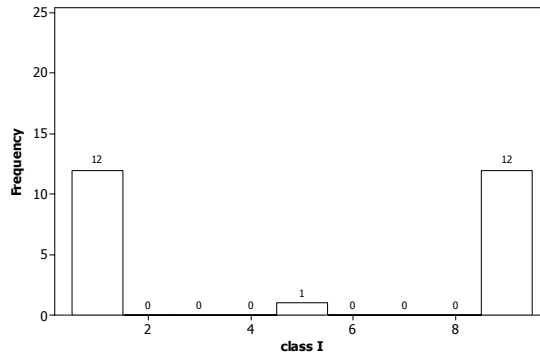
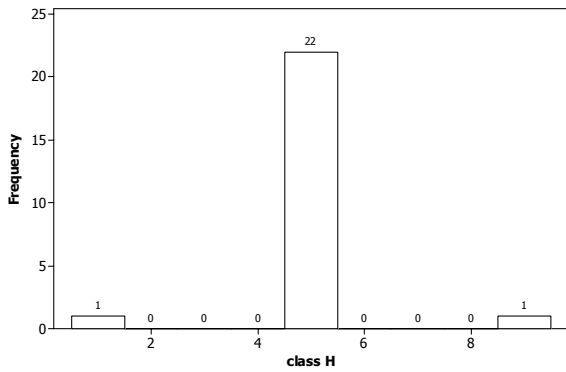
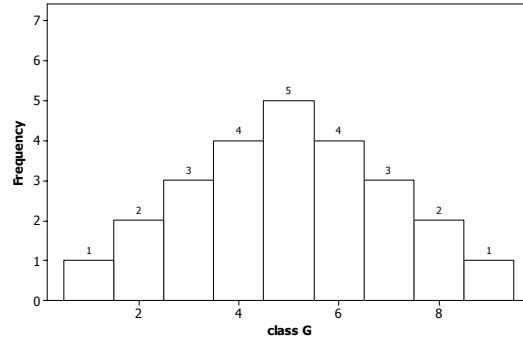
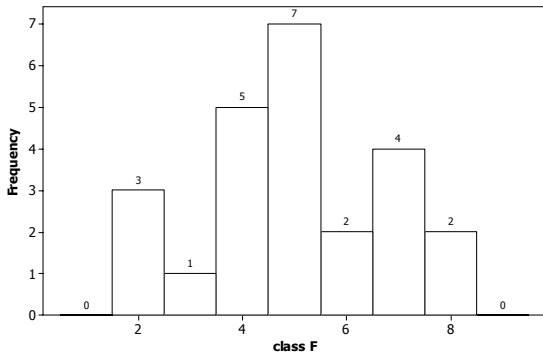
(u) Does the closeness of these medians indicate that the pamphlets are perfectly matched, or at least well-matched, to the patients' reading levels? Explain.

(v) What proportion of the patients have a reading level below that of the easiest-to-read pamphlet?

This example is meant to remind you that center is only one aspect of a distribution; we are often more interested in considering the entire distribution.

Example: Hypothetical Quiz Scores

Consider the distributions of quiz scores in the following five classes:



(w) Between classes F and G, which class do you suspect has more variability in its quiz scores? Explain.

(x) Among classes H, I, and J, which class do you suspect has the most variability in its quiz scores? Which do you suspect has the least? Explain.

Most:

Least:

(y) Open the Minitab worksheet `valuesFJ.mtw`. Calculate the standard deviation and IQR of the quiz scores in each class. Record the results below.

Class	F	G	H	I	J
Std. dev.					
IQR					

(z) Re-answer (w) and (x) in light of these calculations.

This example is intended to reveal that bumpiness and variety are not the same as *variability*.

Creating Hypothetical Examples

(aa) Create an example of 10 hypothetical exam scores such that (feel free to use Minitab):

- the standard deviation is as small as possible
- the standard deviation is as large as possible
- the mean is larger than 90% of the scores
- the mean is larger than twice the median
- the standard deviation is positive, the IQR equals zero, and the mean is greater than twice the median
- Explain how it can happen when a person moves from state A to state B, the mean IQ in both states could decrease.